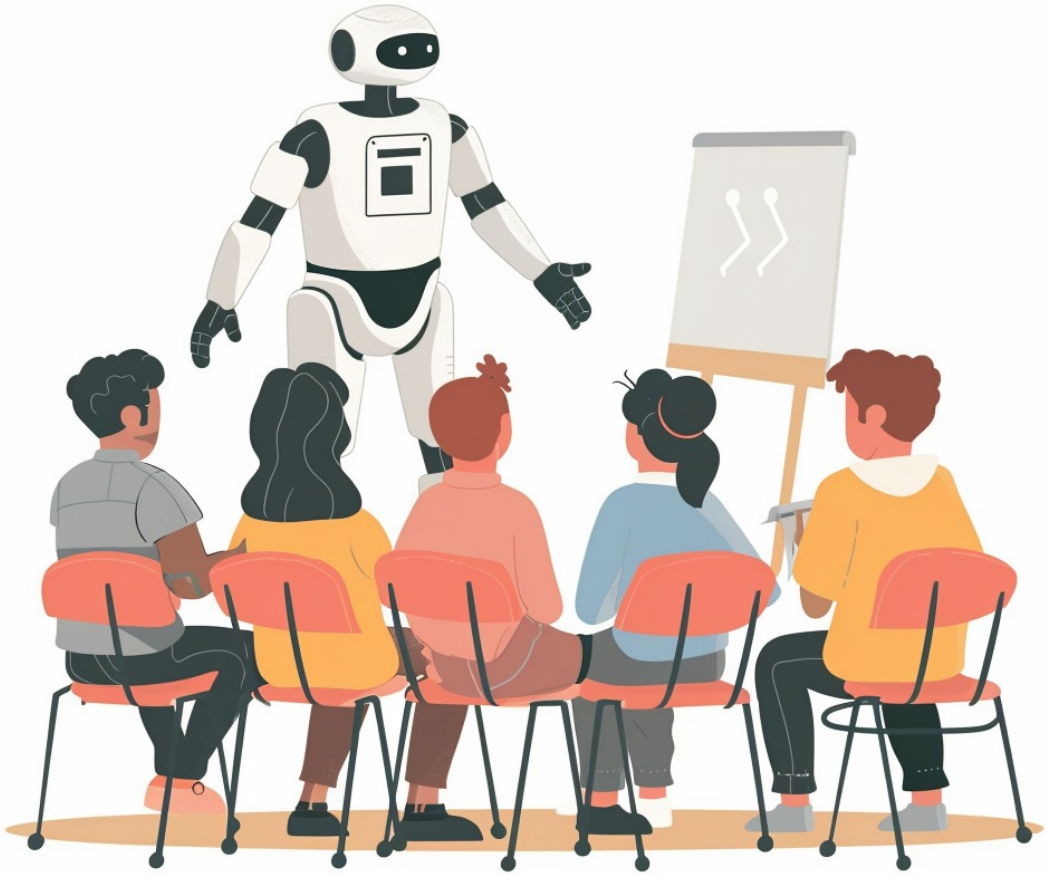


Two Sample Studies

(+ a few first words on the user study “mindset”)

Ulrike Kuhl

Bielefeld University, Bielefeld, Germany

Designing a User Study

To what end do you want to test **what** with **whom** in which **context**?

Adadi & Berrada (2018) IEEE Access

Hoffman et al. (2023) Kunstl. Intell.

Designing a User Study

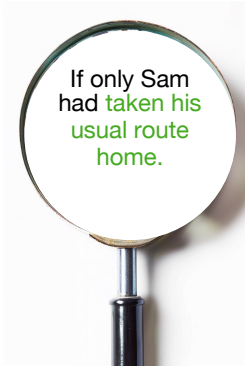
To what end do you want to test **what** with **whom** in which **context**?

Purpose/Objective: Examining…

…humans

…models

…human-AI interaction



…explanations

Adadi & Berrada (2018) IEEE Access

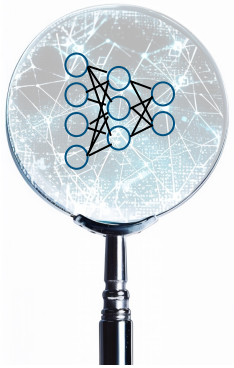
Hoffman et al. (2023) Kunstl. Intell.

Designing a User Study

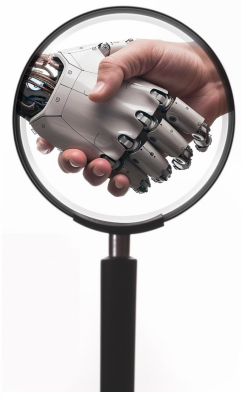
To what end do you want to test **what** with **whom** in which **context**?

Purpose/Objective: Examining…

…humans



…models



…human-AI interaction

…explanations

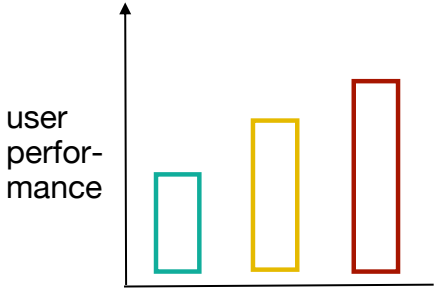
Focus/Variables - examples:

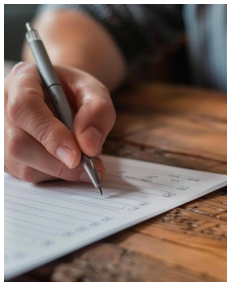
If the odor had been less foul, the food would have been deemed edible.

vs.

Lime (Ribeiro et al., 2016, ACM SIGKDD)

--> compare XAI methods with each other





--> assess performance, subjective measures, etc.

Adadi & Berrada (2018) IEEE Access

Hoffman et al. (2023) Kunstl. Intell.

Designing a User Study

To what end do you want to test **what** with **whom** in which **context**?

Target User Group:


vs.

vs.


--> Challenges: sample size, recruitment, inclusion / exclusion criteria

Adadi & Berrada (2018) IEEE Access

Hoffman et al. (2023) Kunstl. Intell.

Designing a User Study

To what end do you want to test **what** with **whom** in which **context**?

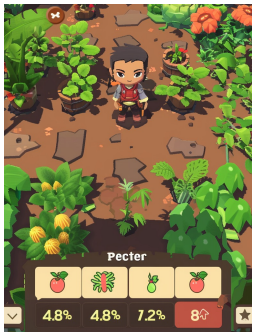
Target User Group:


vs.

vs.


--> sample size, recruitment, inclusion / exclusion criteria

Application / Environment:


vs.


vs. „Imagine you're a gardener...“


Med-AI detected cancer. Do you agree?

--> User's world knowledge matters!

Adadi & Berrada (2018) IEEE Access

Hoffman et al. (2023) Kunstl. Intell.

What has been done: Sample Study I

Motivation

To what end do you want to test **what** with **whom** in which **context**?

To what end?

James	
Gender	Male
Weight	81kg

categorical

continuous

Examine impact of continuous and categorical features in counterfactual and causal explanations

"if only..." "because..."

Warren, Byrne & Keane (2023) ACM IUI

What has been done: Sample Study I

To what end do you want to test **what** with **whom** in which **context**?

To what end?

James	
Gender	Male
Weight	81kg

categorical

continuous

Examine impact of continuous and categorical features in counterfactual and causal explanations

"if only..." "because..."

What?

Human prediction accuracy w.r.t. a system's decisions + subjective satisfaction & trust in the system

Warren, Byrne & Keane (2023) ACM IUI

Warren, Byrne & Keane (2023) ACM IUI

Motivation

What has been done: Sample Study I

To what end do you want to test **what** with **whom** in which **context**?

Whom?	"Representative" human users

Prolific

Warren, Byrne & Keane (2023) ACM IUI

Motivation

What has been done: Sample Study I

To what end do you want to test **what** with **whom** in which **context**?

Whom? "Representative" human users

Context? *SafeLimit* App

James	
Gender	Male
Weight	81kg
Units	6
Duration	105 mins
Stomach	Full
Limit	**?**

Over the limit ○ Don't know ○ Under the limit ○

Warren, Byrne & Keane (2023) ACM IUI

What has been done: Sample Study I

Motivation

To what end do you want to test **what** with **whom** in which **context**?

Whom? "Representative" human users

Context? *SafeLimit* App

James	
Gender	Male
Weight	81kg
Units	6
Duration	105 mins
Stomach	Full
Limit	**?**

Over the limit ○ Don't know ○ Under the limit ○

Counterfactual

Duration	105 mins
Stomach	Full
Limit	Over

Explanation
If James had drunk **5 units** instead of **6 units**, he would have been **under the limit**.

✓ Over the limit ◉ Don't know ○ Under the limit ○

Causal

Duration	105 mins
Stomach	Full
Limit	Over

Explanation
James is **over the limit** because he drank **6 units**.

✓ Over the limit ◉ Don't know ○ Under the limit ○

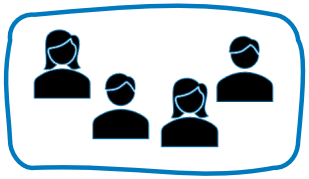
What has been done: Sample Study I

Design

Between-Subjects

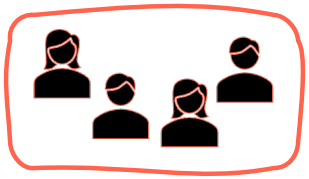
--> 3 groups in 3 conditions

Group 1: Counterfactual



vs.

Group 2: Causal



vs.

Group 3: Control

Explanation
If James had drunk **5 units** instead of **6 units**, he would have been **under the limit.**

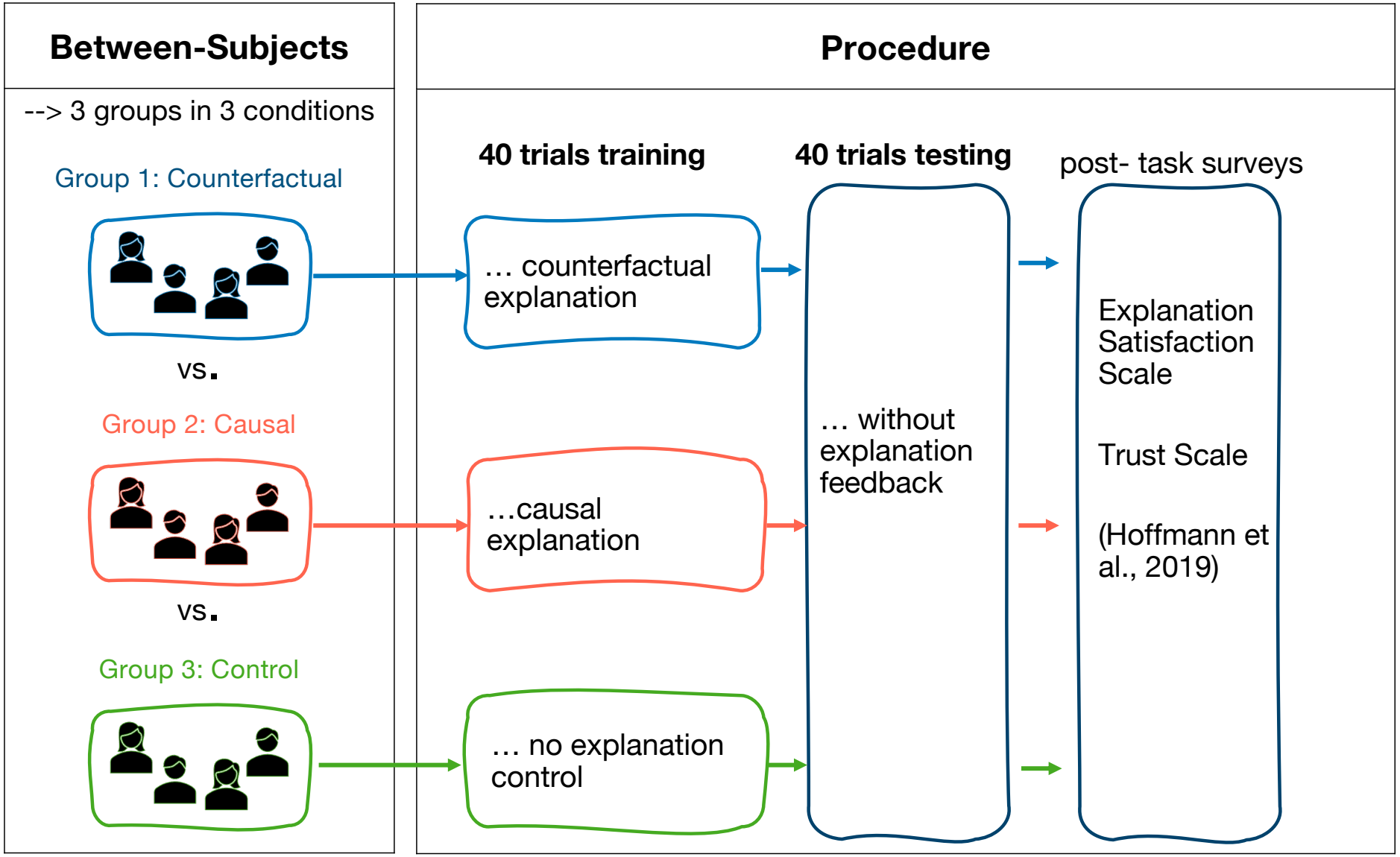
Explanation
James is **over the limit** because he drank **6 units.**

Explanation
James is **over the limit.**

Warren, Byrne & Keane (2023) ACM IUI

What has been done: Sample Study I

Materials & Procedure

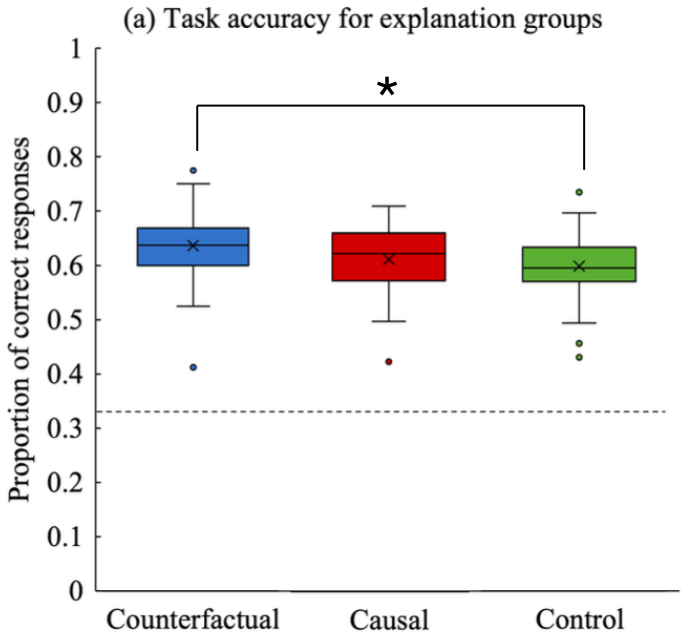


Warren, Byrne & Keane (2023) ACM IUI

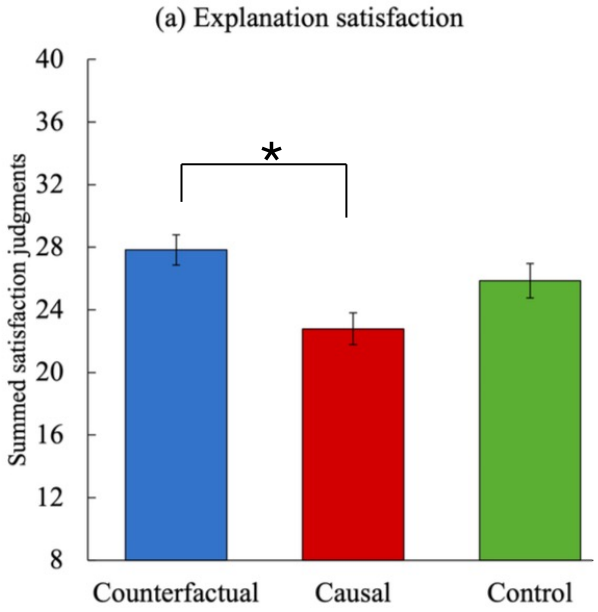
What has been done: Sample Study I

Results

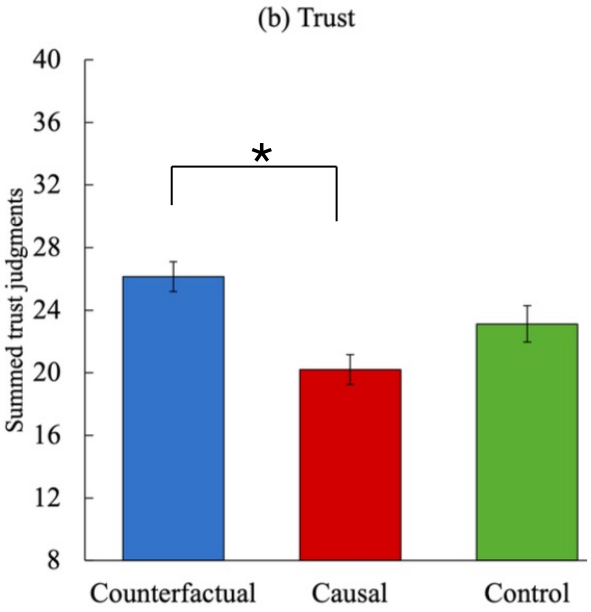
How accurate?



How satisfied?



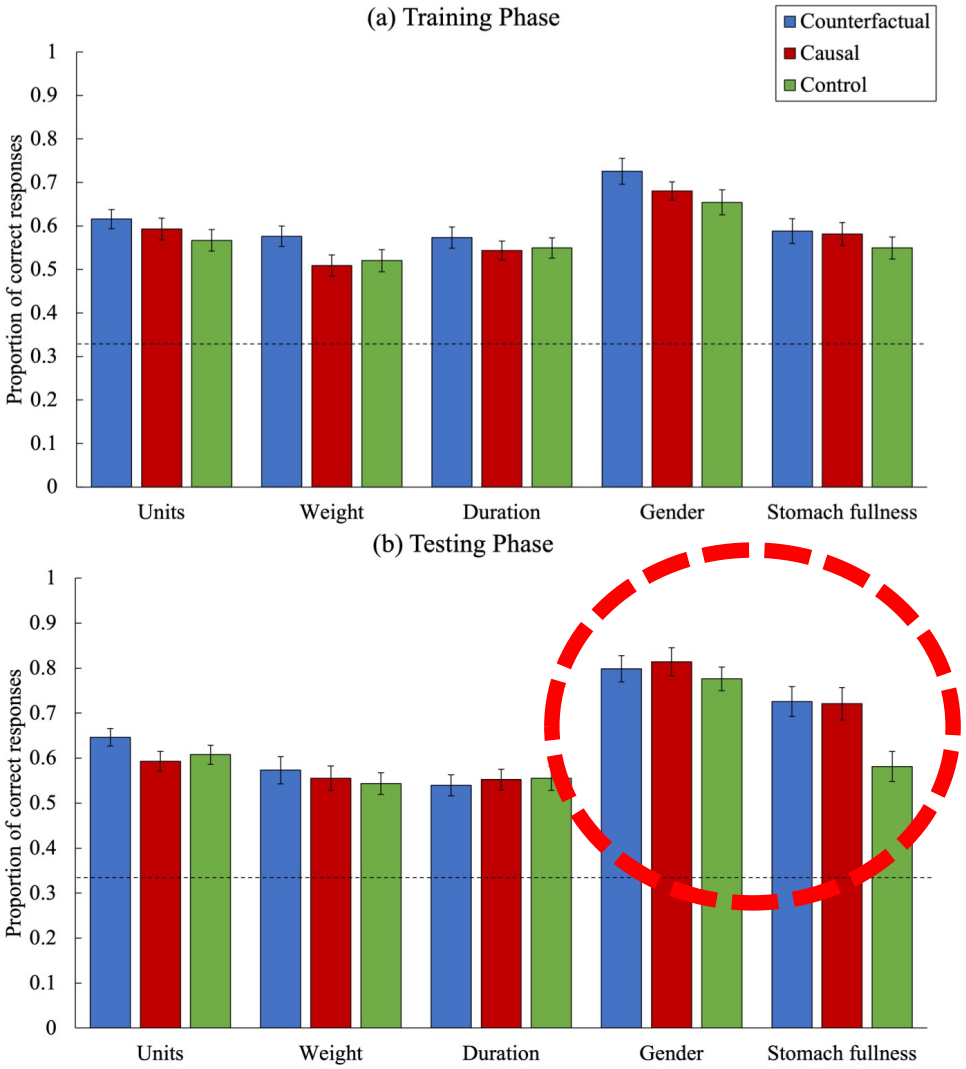
How trusting?



[Warren, Byrne & Keane (2023) ACM IUI]

What has been done: Sample Study I

Results



→ accuracy improved from training to test for categorical, but not for continuous features

Warren, Byrne & Keane (2023) ACM IUI

Warren, Byrne & Keane (2023) ACM IUI

What has been done: Sample Study I

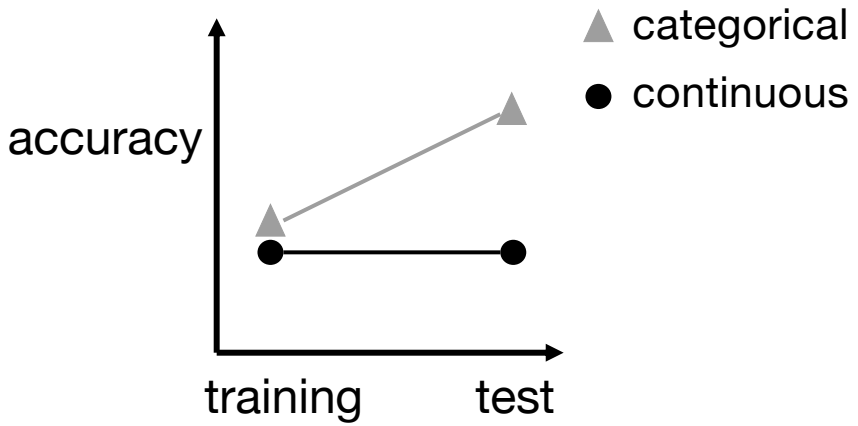
Results

To what end? Examine impact of continuous and categorical features in counterfactual and causal explanations

Accuracy Improvement:

- generally:
  explanations > no explanations
- specifically:
  CF > no explanations

Feature Types:



Subjective Evaluations (satisfaction and trust):

- CF > causal

→ explanation type and fundamental aspects of features matter!

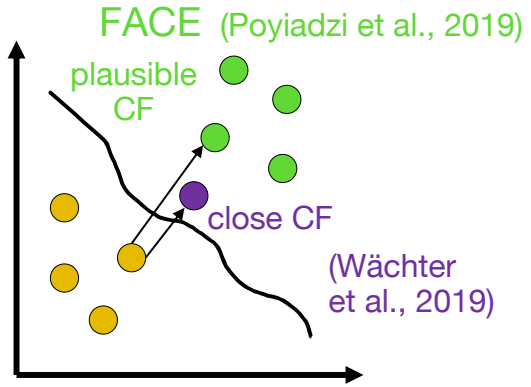
What has been done: Sample Study II

Motivation

To what end do you want to test **what** with **whom** in which **context**?

To what end?

Examine usefulness of a plausibility constraint for generation of counterfactual explanations



Kuhl, Artelt & Hammer (2022) FAccT

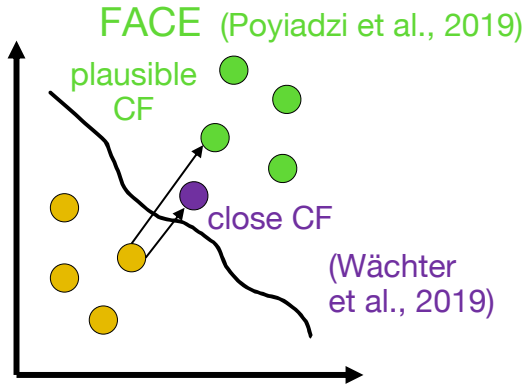
What has been done: Sample Study II

Motivation

To what end do you want to test **what** with **whom** in which **context**?

To what end?

Examine usefulness of a plausibility constraint for generation of counterfactual explanations



What?

Human learning of an unknown system, supported by explanations

Kuhl, Artelt & Hammer (2022) FAccT

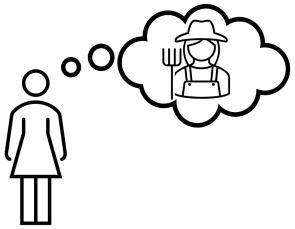
What has been done: Sample Study II

To what end do you want to test **what** with **whom** in which **context**?

Whom? "Representative" human users

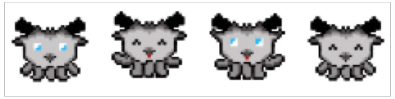
Context?

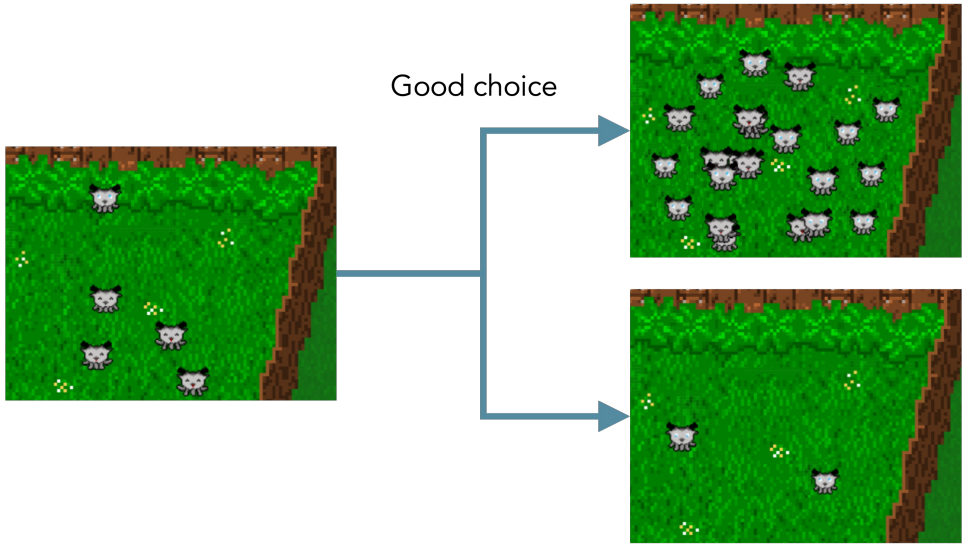
Abstract scenario:
Alien Zoo framework



… selects plants…

… to feed aliens:





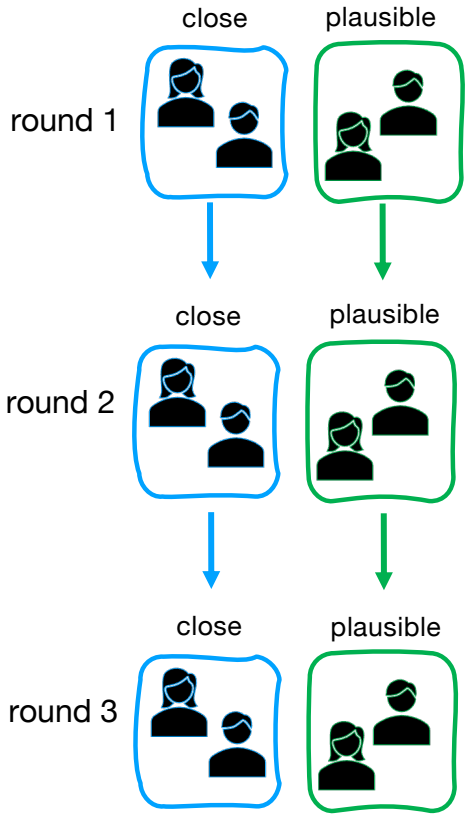
Kuhl, Artelt & Hammer (2022) FAccT

What has been done: Sample Study II

Design

Within-Subjects, repeated measures

--> 2 groups in 2 conditions, repeatedly measured



Kuhl, Artelt & Hammer (2022) FAccT

What has been done: Sample Study II

Materials & Procedure

Within-Subjects, repeated measures

--> 2 groups in 2 conditions, repeatedly measured

Group 1: close

Group 2: plausible

Procedure

Play the game, interleaved with explanations

rounds 1 & 2

close CFE

plausible CFE

rounds 3 & 4

close CFE

plausible CFE

rounds 11&12

post- task surveys

Adapted system causability scale

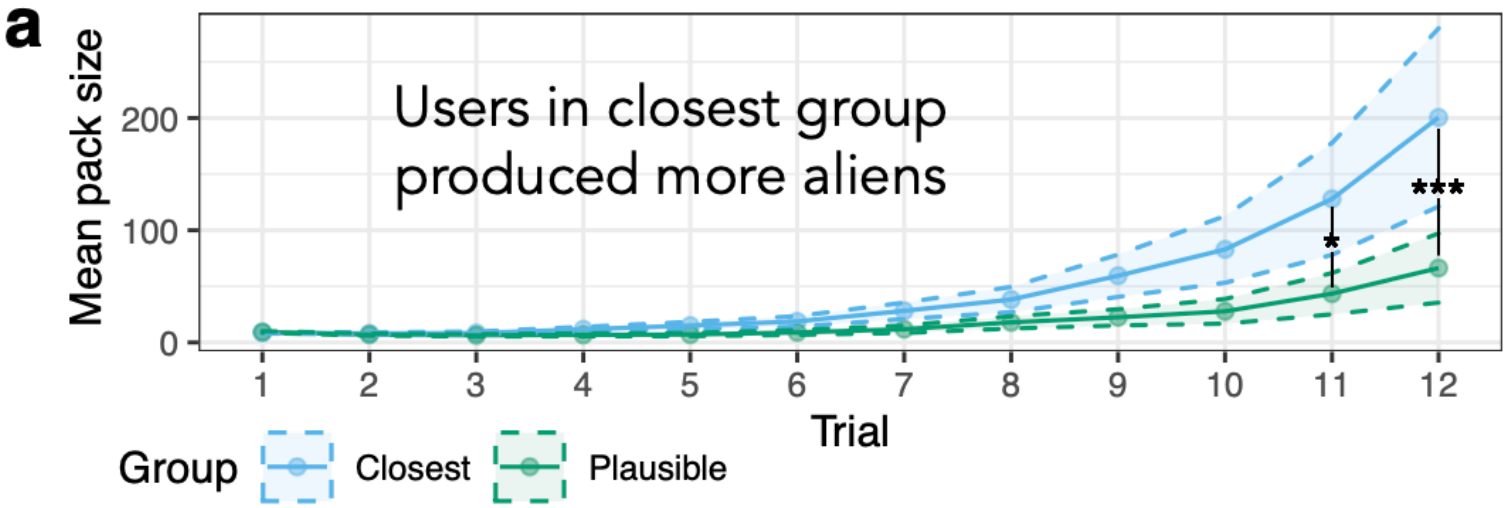
(Holzinger et al., 2020)

time

Kuhl, Artelt & Hammer (2022) FAccT

What has been done: Sample Study II

Results



Yet:

- participants in both groups could distinguish important from unimportant features
- no differences in subjective measures were detected

Kuhl, Artelt & Hammer (2022) FAccT

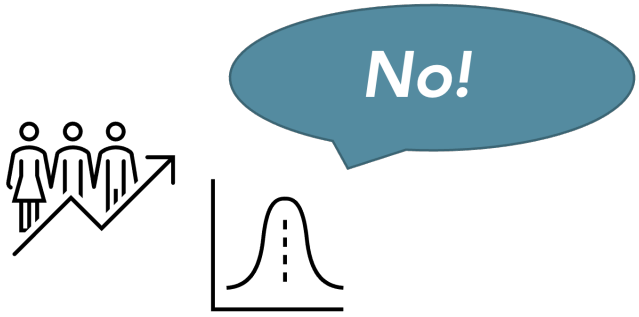
What has been done: Sample Study II

Results

To what end?

Examine usefulness of a plausibility constraint for generation of counterfactual explanations

Do ***novice users*** benefit from a plausibility contraint on computing CFEs, when tasked to ***gain new information from an unknown system*** in an ***abstract domain***?



Staying more similar to the original instance is more effective for learning, if no prior knowledge exists!

Kuhl, Artelt & Hammer (2022) FAccT

Plan for the Day

CF Tutorial

TIME	Topics
9:00 AM	***Introduction***
	Hello and Introducing Ourselves!
	Hands-on: *Trying Our Study (follow link)*
9:30 AM	***Historical Fundamentals of Counterfactuals***
	From Philosophy to XAI (via Psychology)
	Two Sample User Studies and Q&A
10:30 AM	COFFEE (10:30-11:00)
11:00 AM	***Fundamentals of Counterfactuals in AI***
	Formalisation
	Modelling Approaches & Key Constraints
11:30 AM	***Using Counterfactual Algorithms***
	Hands-on: *A Counterfactual Toolbox (AA)*
	Hands-on: *Checking Out Notebooks and Q&A*
12:00 PM	***Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design
12:30 PM	LUNCH (12:30-14:00)
2:00 PM	***Algorithmic Growth Points***
	Computational Future Directions and Q&A
2:30 PM	***More Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design (cont.)
3:00 PM	COFFEE (15:00-15:30)
3:30 PM	***From Fundamentals to an Actual User Study***
	User Studies II: A More Complex Design
	User Studies III: Even More Complex Designs
	Hands-on: *Looking At Our Study*
5:00 PM	***Closing Session, Discussion and Final Q&A***
	TUTORIAL END